# Does the socioeconomic status and demography of a country influence the success of a YouTube channel?

**Group Members: Qianwen Luo; Jiaxin Guo; Tingjun Lin**

## Introduction

The emergence of content creation platforms, especially YouTube, as a leading channel for distributing a wide range of content to consumers worldwide, has greatly changed the digital environment. Even though making YouTube videos is profitable is well-acknowledged, how different types of channels affect earnings is still understudied due to the lack of data. YouTube has not only become a significant source of entertainment and information but also a substantial avenue source. This research aims to explore an intriguing aspect of YouTube's global influence: the impact of socioeconomic and demographic factors of a country on the success metrics of YouTube channels.

The proliferation of digital media has enabled individuals from diverse geographic and socioeconomic backgrounds to reach global audiences, especially in developing countries. However, the extent to which the socioeconomic and demographic profile of a country influences the success of its YouTube channels remains an underexplored area. Broadly speaking, does the prosperity of digital media bridge or widen the gap between developed and developing countries? Specifically, given the same quality of content, will the location of YouTube channels hinder its profits? On the one hand, conceptualizing and measuring the socioeconomic status of different countries is a complex task. On the other hand, whether variables about socioeconomic status and demography have collinearity issues is still at the center of debate. Furthermore, whether the casualty between these two variables influences our inference on the success of YouTube channels is unanswered by the previous research. We define the success of YouTube channels as multidimensional for the purpose of our study. Based on the quality of our data, success on YouTube can be quantified through various metrics, including subscriber count, video views, and estimated earnings. These metrics not only reflect the popularity and reach of a channel but also its economic viability and influence. Additionally,

YouTube's global reach allows content creators from diverse locations to cultivate audiences and generate revenue. Analyzing the relationship between the socioeconomic and demographic profile of a country and the subsequent account success will examine whether factors like a country's unemployment rate, population size, and other attributes impact the earnings potential for creators based in specific geographical locations.

The intersection of digital media success with socioeconomic and demographic variables presents a unique opportunity to understand the digital landscape from a new perspective. It can offer critical insights into the varying monetization models and audience dynamics that underpin success in specific areas.

This analysis has the potential to inform content creators, digital marketers, policymakers, and platform stakeholders about the most effective strategies for optimizing revenue generation and fostering sustainable growth within their respective content categories. For instance, insights from such research can guide policymakers in understanding the impact of digital infrastructure and education systems on digital content creation. YouTubers can use the insights to tailor their content strategy according to the demographic and socioeconomic contexts of their target audiences.

## Data Preparation

To understand the influence of a country's socioeconomic and demographic profile on the success metrics of YouTube channels, we looked for relatively updated datasets that contain variables including revenues, geographical elements, content categories, etc. We first found a data list from HypeAuditor, which lists the Top 1,000 YouTube channels and their follower amounts, views, countries of origin, ranking, usernames, content categories, accumulated visits, likes, comments, etc. However, such a list of data only has around 1000 rows. To perform a more robust analysis with a larger dataset, we incorporated it with a large dataset called “Trending YouTube Video Statistics”, integrating both video data and demographic and socioeconomic information.

In preprocessing data, we removed all the “N/A” numbers and duplicated rows. `dplyr` and `readr` libraries were used for data manipulation and reading. In merging all the CSV files to one large dataset, we extracted the country name from each file name, appended the corresponding country name as a new column, and combined all individual datasets into a single large dataset, `trending_ytb`. We filtered out videos with disabled comments, disabled ratings, or those that were removed due to errors. Next, we simplified country codes to their first two letters, then mapped these codes to full country names (e.g., "US" to "USA", "GB" to "Great Britain"). We selected relevant columns like `video_id`, `trending_date`, `channel_title`, `category_id`, `views`, `likes`, `dislikes`, `comment_count`, and `country`.

To integrate the data with demographic and socioeconomic variables like tertiary education enrollment, population, unemployment rate, urban population, latitude, and longitude. We merged this demographic data with the YouTube trending video data based on the country. Finally, we removed duplicate entries and missing values to refine the dataset further.

Specifically, we created a new variable called “success_index”, which is a calculated metric that provides a standardized way to assess the overall engagement or popularity of each YouTube video in your dataset, considering views, likes, and comment counts.

This process resulted in a large, clean dataset that combines detailed statistics of trending YouTube videos. This enriched dataset can provide insights into how the socioeconomic and demographic profile of a country influences the success metrics of YouTube channels in that country. Among the 15 variables in the merged dataset, we mainly categorize them into two groups. The first category is variables that are used to measure the success of a YouTube video. The other category includes variables that contain demographic and socioeconomic information about the YouTube video. Together, these variables allow an analysis of influencer reach, audience preferences, engagement levels, and geographic trends across top YouTube creators producing streaming video content. The data facilitates a quantitative evaluation of streaming success factors and a comparison between different countries and factors.

## Method

a. Descriptive statistics

To get an initial, rough understanding of views data across demographic/socioeconomic factors, we plotted histograms of the logarithm of the success indices, grouped by category variables that we created. This step is aimed to see whether there are any obvious trends among different levels of demographic/socioeconomic variables(unemployment rate, urbanization and territory education level) in views.

During the process, we noticed that there are non-finite values (like NaN or Inf) in your data, particularly in the success_index column or as a result of the log(success_index) transformation. This can happen if success_index contains zero or negative values, as the logarithm of zero or a negative number is undefined in real numbers. We filtered them out or transform them in a way that avoids NaN values.

b. Regression analysis

i. *log(Success_index) ~ log(Population) + Education + Unemployment_rate + log(Urban_population)*

ii. *log(Success_index) ~ log(Population) + Education + Unemployment_rate*

To have more insights into the relationship between the success of YouTube channels and demographic and (or) socioeconomic factors, we perform some regression analysis and test whether the relationship, if any, is significant.

c. Smoothing Method: KDE and LOESS

In order to estimate the probability density function of the "success_index" variable that we created, so as to lay a foundation for the analysis next, we specified kernel = 'gaussian', which means that the KDE used a rectangular kernel for smoothing. Gaussian kernel is the most commonly used kernel because it produces a smooth, bell-shaped curve and has nice mathematical properties. It is suitable for a smooth, general overview of the data distribution.

$$K(t) = (1/\sqrt{2\pi})\, e^{\left(-\frac{t^2}{2}\right)} (t \in (-\infty, \infty))$$

Besides estimating the pdf of the "success_index" variable, we also wanted to visualize general trend of the variable across time. Treating it as time-series data, we chose to use Locally Estimated Scatterplot Smoothing to visualize its trend. We chose LOESS because we noticed that the data is quite noisy, as it is often impacted by a number of factors. It also follows a non-linear trend(from method b, we saw that there really isn't a clear linear trend between our explanatory variables and success_index), suggesting that it cannot be easily captured with a simple linear model.

In general, LOESS uses a polynomial to fit a subset of the data at a time. By default, the weight function that we used for smoothing is the tri-cube weight function:

$$w(d) = (1 - |d|^3)^3,$$

where d is the distance of a given data point on the curve being fitted, scaled to lie from 0 to 1.

It is worth noting that LOESS has locality property. This means that the algorithm captures local patterns in the data. LOESS involves an iterative procedure where a series of weighted least squares regressions are performed. At each point in the dataset, a local regression is fitted to predict the value of the response variable for that point. This results in a smooth curve that runs through the dataset, revealing the underlying trend.

d. Bootstrap Confidence Intervals

We also want to perform a stratified bootstrap analysis to see views data for each country. This would prepare us as we can take a first look at how our data look like across countries–which is our final goal of this research. What we did is to loop over each unique country in merged_data, subset the data for that country, and then perform bootstrap analysis using 1000 replications.

For each country, two types of confidence intervals are presented:

- Basic Confidence Interval (Basic_CI): uses the bootstrap distribution and adjusts it based on the difference between the bootstrap mean and the observed mean. This interval assumes a roughly normal distribution of the mean.
- Percentile Confidence Interval (Percentile_CI): takes percentiles from the bootstrap distribution. (often the 2.5th and 97.5th for a 95% CI).

e. Anova and Permutation tests

One of our datasets, titled "Global YouTube Statistics," offers a comprehensive overview of the top YouTube channels, encompassing a wide range of metrics such as subscriber counts, video views, upload frequency, and, crucially, earnings data. It provides detailed information on the performance of popular YouTube channels across multiple countries and categories. Key variables such as the highest yearly earnings allow us to quantify video success and revenue generated on the platform.

The ANOVA method is well-suited for this comparison as it statistically tests if the mean earnings differ between various countries and categories. A significant ANOVA F-test suggests location and employment type impact revenue potential on YouTube. We can further validate these findings using permutation testing. By randomly shuffling the country and category labels, we can simulate the null hypothesis of no true differences between groups. Comparing the real observation to distributions under this null hypothesis provides a non-parametric test of significance. These methods provided insights into whether the differences observed in earnings based on country and employment category are statistically significant or merely products of random variation. The country and unemployment category associated with each video enables comparative analysis across these groupings. The country associated with each channel enables comparative analysis across these groupings. For example, we can analyze whether significant differences exist in earnings between channels from the United States versus India or between different levels of unemployment categories.

i. Analysis of Variance

- In this case, the results of the ANOVA tests will reveal whether there are statistically significant differences in the *highest yearly earnings* across countries and unemployment categories.
- Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It is particularly useful when comparing the means of three or more independent groups to determine if there is any statistically significant difference between them.

The formula for the one-way ANOVA F-test statistic is:

$$F = \frac{MS_{between}}{MS_{within}}$$

where $MS_{between}$ $and$ $MS_{within}$ are Mean Squares between Groups and Mean Squares within Group. They are given by:

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

Here, SSbetween (Sum of Squares Between Groups) and SSwithin (Sum of Squares Within Groups) are calculated as follows:

$$\text{SSbetween} = \sum_{i=1}^{k} n_i(\overline{X_i} - \overline{X})^2$$

$$\text{SSwithin} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{X_{ij}} - \overline{X_i})^2$$

where k is the number of groups, ni is the number of observations in group i, $\overline{X_i}$ is the mean of group i, $\overline{X_{ij}}$ is the jth observation in group i, and $\overline{X}$ is the overall mean.

ii. Permutation Test

- The permutation test is a non-parametric method used to determine the significance of the difference between two groups. This test involves repeatedly shuffling the data and recalculating the test statistic to create a distribution of the statistic under the null hypothesis.
- Permutation tests will be applied to compare the *highest yearly earnings* between specific countries or levels of unemployment categories. This test will help determine if any observed differences are statistically significant or could have occurred by chance.
- The test statistic used is the difference in means between the two groups. The process is as follows:
    - - Combine the data from both groups into a single dataset.
    - - For a large number of permutations (e.g., 1000, 10000):

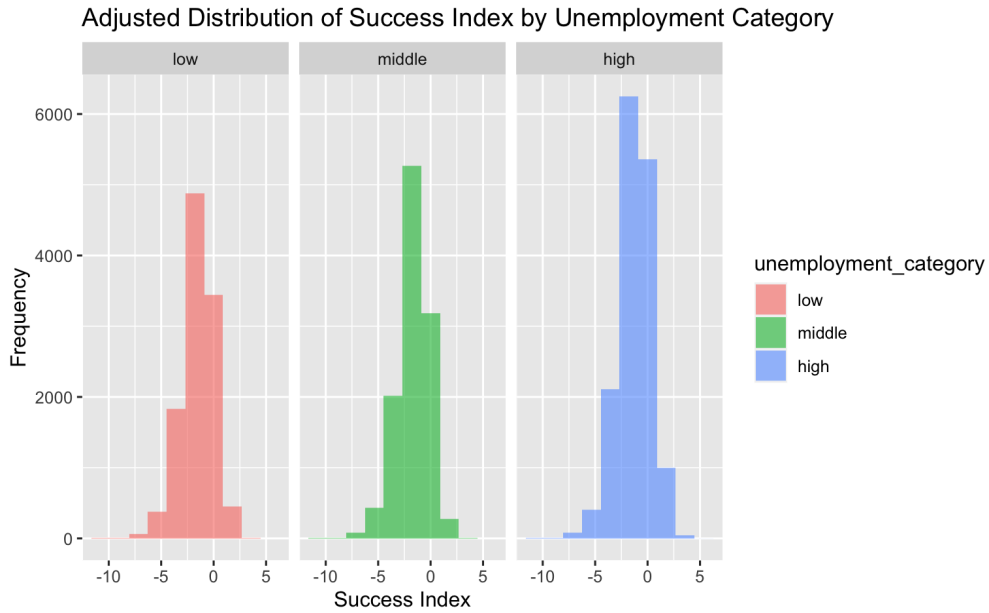
- - Randomly shuffle the combined dataset.
- - Split the shuffled dataset into two new groups, maintaining the original group sizes.
- - Calculate the difference in means between these two new groups.

- - Calculate the proportion of permutations where the difference in means is greater than the observed difference. This proportion is the p-value.

## Simulations & Analysis

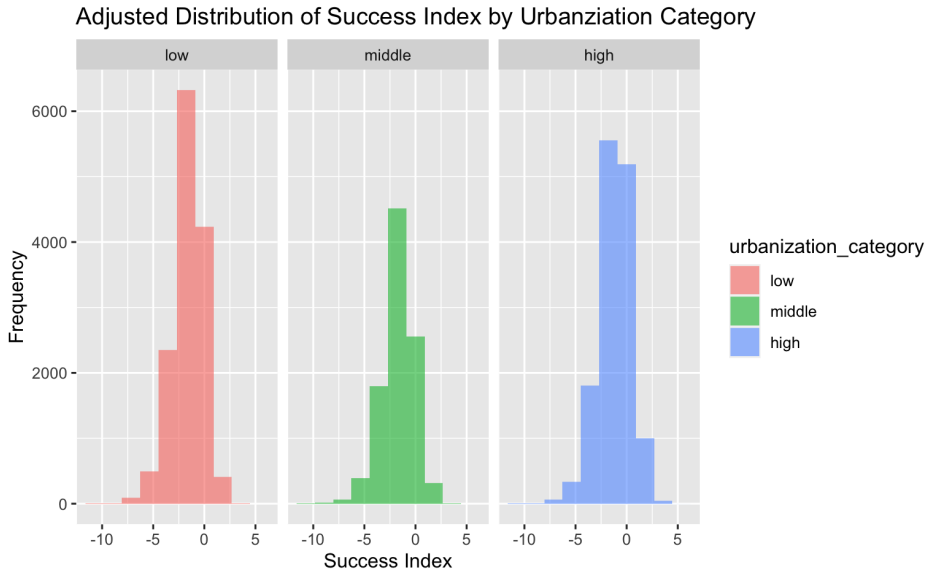
a. Histograms

The first histogram graph is grouped by “unemployment_category” to high, middle, and low levels. High means the unemployment rate of the origin place of this YouTube video is higher than the median unemployment rate, and low means the unemployment rate of the origin place of this YouTube video is lower than the median unemployment rate.

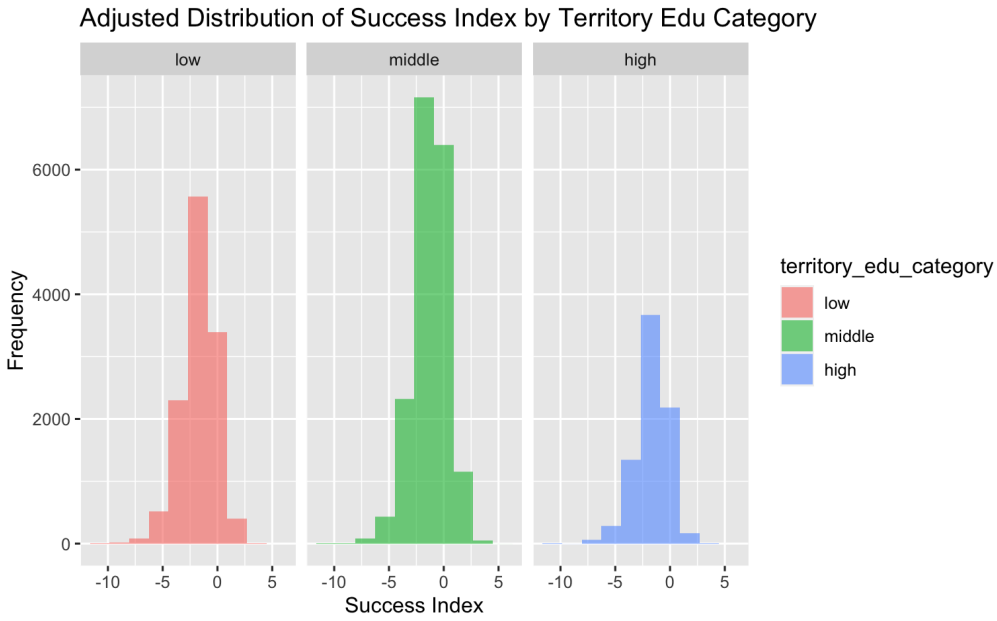



Looking at the shape, spread, and central tendency of the Success Index distribution within each unemployment category, we can see that the 'high' unemployment category might have a slightly wider distribution, suggesting more variability in the Success Index within this category. We observed that the graph for a high level of unemployment category is slightly left-skewed,

indicating that while most videos have a relatively high success index, there are a few videos with significantly lower success.




The highest frequency in the central part of the histogram for the low urbanization category indicates that most videos from less urbanized areas have a success index around the median. This could suggest a consistent pattern of audience engagement in these areas.



The highest frequency in the central part of the histogram for the middle education category suggests that most videos from territories with a middle level of education have a moderate success index. This could indicate a balanced audience engagement in these territories. The lower frequency in the central part for the high education category might suggest less concentration around the average success index.

b. Regression Result

Main Regression model:

$$log(Success\_index) \sim log(Population) + Education + Unemployment\_rate + log(Urban\_population)$$

We first do the log transformation for our population and urban population variables. Log transformation for the data like population can help us standardize the variable to avoid skewness and the impact of extreme values. Applying a log transformation can linearize relationships that are inherently exponential. For example, if the impact of population on the dependent variable is multiplicative rather than additive, taking the log may make the relationship more linear. The results of our regression are shown in the following table:

| **Variable** | **Coefficient** | **P-value** |
|---|---|---|
| Intercept | 7.12 | 0.00 |
| log(Population) | 0.76 | 0.00 |
| Education | 0.00 | 0.00 |
| Unemployment | -0.03 | 0.00 |
| log(Urban_population) | -1.15 | 0.00 |

We thought that education level may positively correlate with the success of YouTube Channels. Creativity and innovation in content creation may also be influenced by education, with more educated creators potentially producing better material. Access to resources, including

technology and production skills, may be higher among more educated creators, contributing to the production of higher-quality content. However, the coefficient for education is negative and the corresponding p-value for education is less than 0.05, indicating that the relationship is negative. We should also notice that the estimated coefficient is extremely small, casting doubt on the practical significance of interpreting
education in this relationship. In addition, such issues can be caused by the transformation of the success index.

Some might criticize including both population and urban population as our independent variables to study their relationship with success of YouTube channels as they are significantly correlated, and adding urban population does not necessarily improve our model. To address this concern, we perform an ANOVA test for the main regression model and regression model without urban population:

$$log(Success\_index) \sim log(Population) + Education + Unemployment\_rate$$

Based on the F-statistic, the corresponding p-value is less than 0.001. Thus, we can reject the null hypothesis that the main model and the main model without urban population is not significantly different. Surprisingly, we find that the population of a certain country is positively correlated with the success of the channels within that country but urban population is negatively correlated with the success rate. A larger population may provide a larger audience and market for YouTube content creators. In countries with higher populations, there might be more diverse interests and a larger pool of potential viewers, contributing to higher success rates for YouTube channels. On the other hand, higher urban population density might also mean more competition among content creators. In urban areas, there could be a saturation of content, making it more challenging for individual channels to stand out and achieve high success rates.

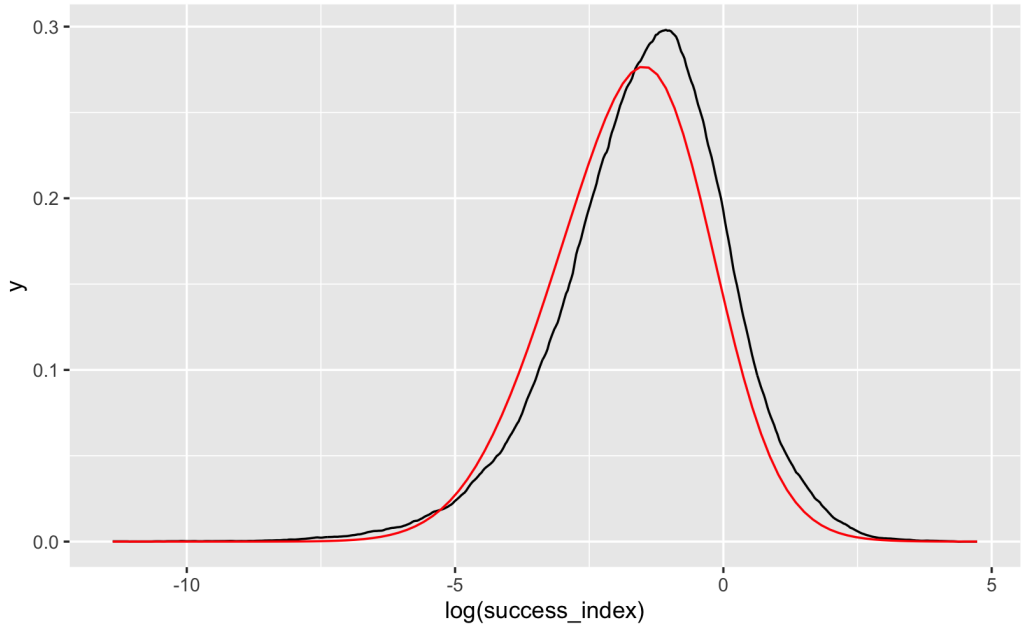
We find a negative correlation between success and unemployment rate. Countries with lower unemployment rates often have more stable and prosperous economies. In such economies, individuals may have higher disposable incomes, leading to increased leisure time and internet usage. This increased free time and spending capacity could contribute to higher viewership and engagement with YouTube content. Countries with lower unemployment rates may also have

better-developed technological infrastructures, including widespread internet access. This infrastructure can facilitate a larger and more engaged audience for YouTube channels.

c. Kernel Density Estimate (KDE)

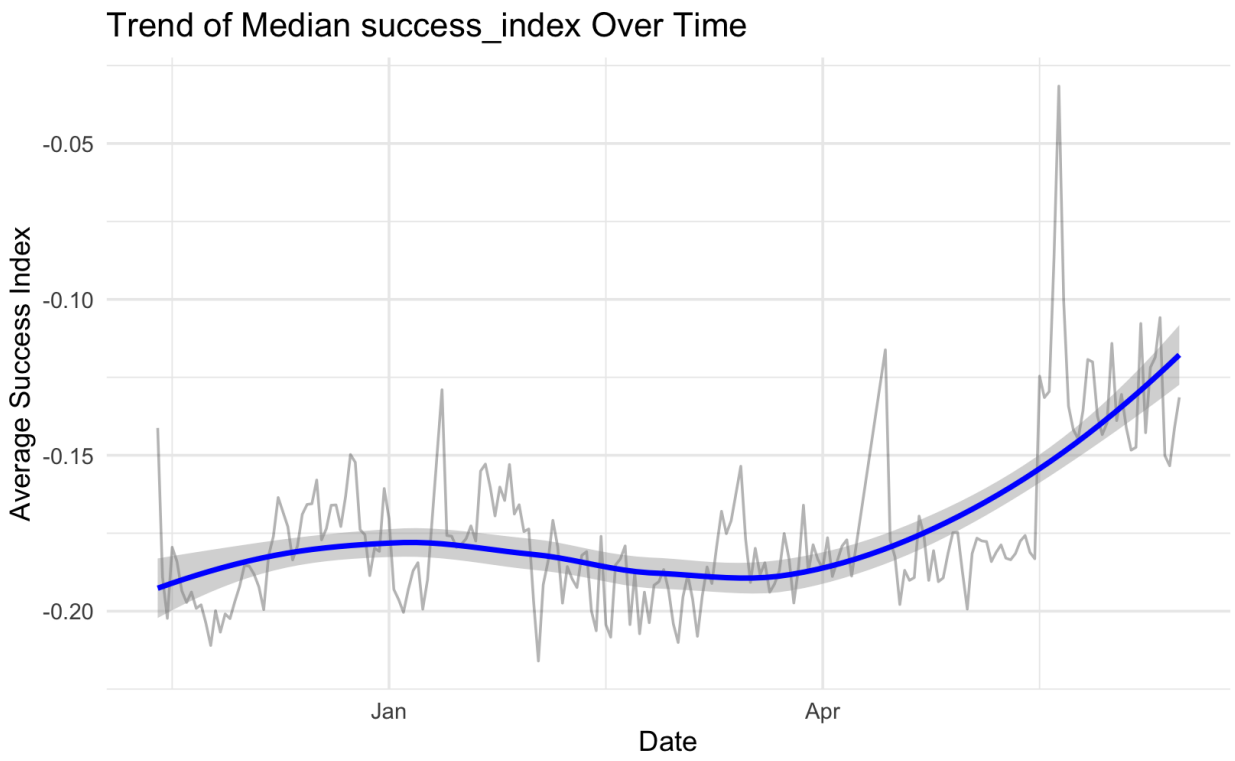
We adjusted the truef function so that it more closely fits the empirical distribution of log(success_index). After parameter-tuning, we eventually chose to use the following "truef" function to simulate the distribution and got a pretty close result, as shown below.

```
truef <- function(x, mean1 = -2, sd1 = 1.5, weight1 = 0.75, mean2 = -1, sd2 = 1, weight2 = 0.25)
{weight1 * dnorm(x, mean = mean1, sd = sd1) + weight2 * dnorm(x, mean = mean2, sd = sd2)}
```




The black line represents the kernel density estimate of the log(success_index). This is an empirical representation of how the data is distributed after a log transformation.

The red line represents the theoretical density function truef, as shown above. The fact that it overlays closely with the empirical density suggests that the log of the success_index fits well with the mixture of the two normal distributions we specified. It might suggest that there are two underlying groups(possibly views and likes) contributing to the observed success index: one that follows a standard normal distribution and another that also follows a normal distribution but has less influence (hence the 25% weight).




The above graph is a result of an algorithm that estimates the latent function in a point-wise fashion. For each value of x, we estimate the value of f(x) by using its neighboring sampled (known) values. LOESS smoothing has been applied to time-series data representing the median success index of YouTube videos over time.

During the simulation, we used the loess function to perform the actual computation for the local polynomial regressions. It computes the fit at a grid of points and then uses these points to interpolate a smooth curve. The local regressions are weighted using a kernel with weights decreasing with distance from the target point.

The raw data is visualized with a light line (geom_line(alpha = 0.3)), which shows all the fluctuations, including noise and outliers. The LOESS smoothed line (geom_smooth(method = "loess", se = TRUE, color = "blue")) overlays this raw data with a smooth, fitted curve that indicates the central trend of the success index over time without being distracted by daily fluctuations. We see that during the period of the data(from 2018 to 2020), there appears a trend of increase-decrease-increase in success_index of a large number of youtube videos.

d. Stratified Bootstrap Confidence Interval

We used stratified bootstrap to calculate the CIs for Youtube videos views statistics aross all the countries, as shown in the following table:

| | | |
|---|---|---|
| $Canada<br>$Canada$Country<br>[1] "Canada" | $Canada$Basic_CI<br>[1] 473239.8 501294.3 | $Canada$Percentile_CI<br>[1] 475250.5 503305.0 |
| $Germany<br>$Germany$Country<br>[1] "Germany" | $Germany$Basic_CI<br>[1] 153332.8 163309.3 | $Germany$Percentile_CI<br>[1] 153705.9 163682.5 |
| $France<br>$France$Country<br>[1] "France" | $France$Basic_CI<br>[1] 77872.25 84942.38 | $France$Percentile_CI<br>[1] 78650.02 85720.16 |
| $India<br>$India$Country<br>[1] "India" | $India$Basic_CI<br>[1] 250488.7 270750.0 | $India$Percentile_CI<br>[1] 251013.8 271275.1 |
| $Japan<br>$Japan$Country<br>[1] "Japan" | $Japan$Basic_CI<br>[1] 102399.2 111033.1 | $Japan$Percentile_CI<br>[1] 102879.7 111513.6 |
| $`South Korea`<br>$`South Korea`$Country<br>[1] "South Korea" | $`South Korea`$Basic_CI<br>[1] 134803.0 149180.2 | $`South Korea`$Percentile_CI<br>[1] 136356.1 150733.4 |
| $Mexico<br>$Mexico$Country<br>[1] "Mexico" | $Mexico$Basic_CI<br>[1] 109175.6 115078.6 | $Mexico$Percentile_CI<br>[1] 109083.3 114986.2 |
| $Russia | $Russia$Basic_CI | $Russia$Percentile_CI |

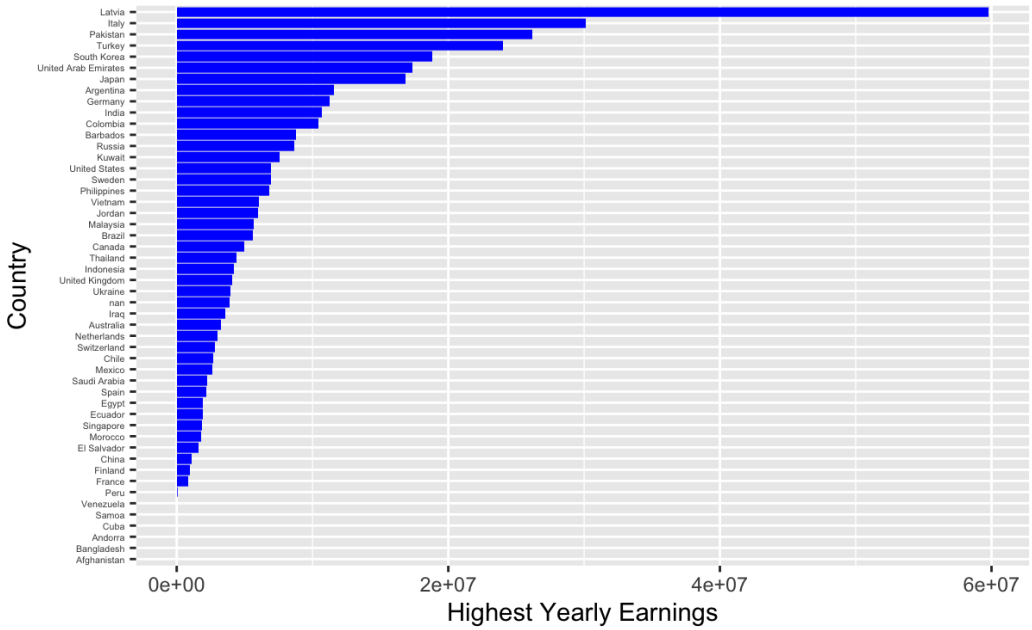
| | | |
|---|---|---|
| $Russia$Country<br>[1] "Russia" | [1] 104019.1 108317.8 | [1] 103987.8 108286.5 |

Some implications can be made based on the above results:

- There is a noticeable variation in YouTube video views across countries, which may be influenced by factors like population size, internet penetration, content preferences, and cultural factors.
- The Basic_CI and Percentile_CI are very similar for each country, which suggests that the underlying distribution of views is approximately symmetric and perhaps normally distributed, as the basic bootstrap method assumes.
- The relatively narrow ranges of the CIs for each country indicate a decent level of precision in the average views estimate, which is beneficial for drawing conclusions about the popularity and reach of YouTube content in different socioeconomic contexts.

e. ANOVA and Permutation Tests

Since we detected different popularities of YouTube videos across countries, we want to start navigating the reasons. To investigate the potential impact of countries' socioeconomic factors on popularities of channels, we can first start by visualizing the YouTube channel's success across different countries and different levels of unemployment categories.

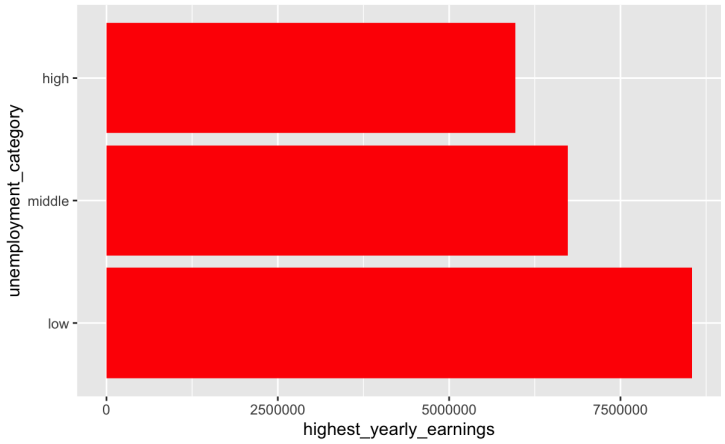



There's a noticeable disparity in the average highest yearly earnings among YouTube channels from different countries, ranging from over $60,000 for channels based in Latvia down to almost zero earnings for some lower-income countries. This suggests that the location of a channel can significantly impact its earning potential.  The top-earning countries might have a combination of factors such as larger audience bases, higher advertising rates, and more lucrative sponsorship opportunities.

Specifically, the highest-earning countries tend to be more developed nations such as Latvia and Italy potentially due to larger audience bases, higher advertising rates, and more lucrative sponsorship opportunities. For example, Latvia's highly educated population, high level of urbanization with widespread internet access enabling content distribution, and integration with European advertising markets likely assist Latvian YouTubers in commanding top earnings on the platform globally.

Conversely, lower-income developing countries face structural disadvantages in maximizing the profitability of YouTube channels. Afghanistan, Bangladesh, Andorra, Cuba, Samoa, and Venezuela registered close to zero average earnings, as indicated by the histogram. Limited internet connectivity, lower viewership demand, and lack of prominent advertiser interest make it

difficult for creators in these geographies to monetize content. Add on political and economic instability, prevalent in countries like Venezuela, leading to even sharper drops in viable revenue sources.




Different levels of unemployment categories on YouTube have varying average highest yearly earnings. The bar representing the low unemployment category has the highest average yearly earnings, exceeding 7,500,000. This suggests that in areas with low unemployment rates, the highest yearly earnings tend to be greater. This could be indicative of stronger economic conditions in these areas, potentially leading to higher disposable incomes and greater spending power. The high unemployment category has the lowest average highest yearly earnings, around 6,000,000. This suggests that areas with higher unemployment rates tend to have a lower average of highest yearly earnings.

The observed trend suggests a correlation between unemployment levels and economic success, as measured by the average highest yearly earnings. Lower unemployment rates might be associated with more robust economic activity, leading to higher earnings. It highlights the influence of broader socio-economic factors on the commercial success of content creators, suggesting a nuanced relationship between employment status and digital engagement.

i. Analysis of Variance (ANOVA)

In this analysis, ANOVA was applied to compare the highest yearly earnings of YouTube channels based on their country of origin. The results are as follows:

```
             Df    Sum Sq   Mean Sq F value   Pr(>F)
Country      49 1.825e+16 3.724e+14   2.058 3.87e-05 ***
Residuals   945 1.710e+17 1.809e+14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-Statistic is a measure of the variance between the group means relative to the variance within the groups. The F-statistic of 2.058 compares the variance between the group means (between different countries) to the variance within the groups (within each country). A higher value indicates a greater difference between the groups so this is a strong indication that the country of origin plays a notable role in influencing the earnings of each video.

The P-value is particularly important. In this case, the P-value is approximately 0.0000387, which is less than the commonly used significance level of 0.05. This suggests that there are statistically significant differences in the highest yearly earnings among YouTube channels from different countries. In other words, the country of origin appears to have a significant impact on the highest yearly earnings of YouTube channels in this dataset. Given the very small p-value, we can reject the null hypothesis that there is no difference in the mean highest earnings across countries.

The ANOVA test has been conducted to compare the highest yearly earnings of YouTube channels across different unemployment categories. The results are as follows:

```
                      Df    Sum Sq   Mean Sq F value Pr(>F)
unemployment_category   2 7.091e+14 3.546e+14   1.811  0.164
Residuals             869 1.701e+17 1.958e+14
```

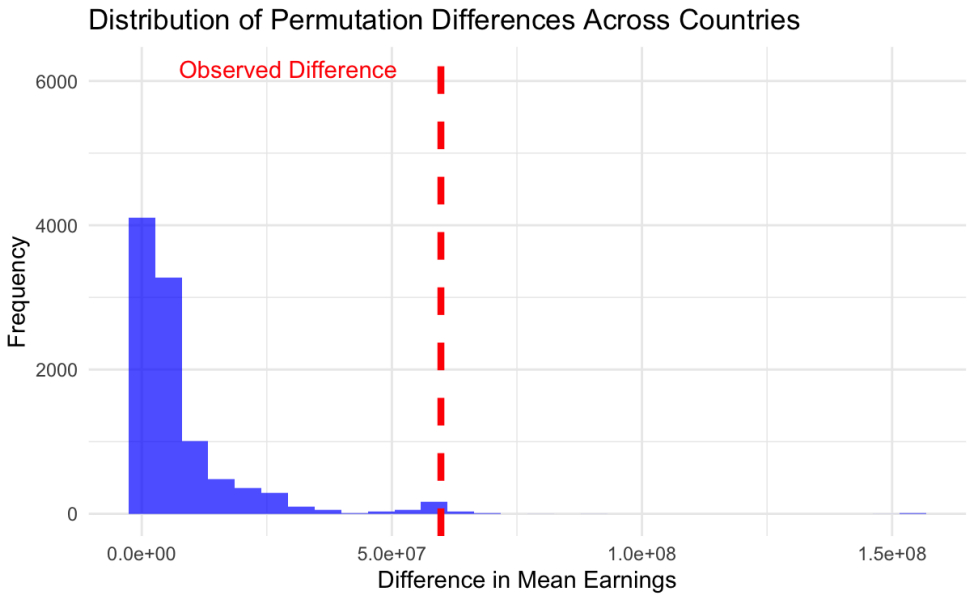
The test has two components with different degrees of freedom (Df). For the 'unemployment_category', Df is 2, indicating three unemployment categories are being compared (low, medium, and high). The F-value of 1.811 suggests the ratio of the variance explained by the

unemployment categories to the variance within the categories (unexplained variance). An F-value greater than 1 suggests that there is more variability between the groups than within them. The p-value of 0.164 is above the conventional alpha level of 0.05. This indicates that the differences in the highest yearly earnings among YouTube channels across different unemployment categories are not statistically significant. In other words, based on this data and this test, we do not have sufficient evidence to conclude that the unemployment category has a significant impact on the highest yearly earnings of YouTube channels. Thus, the ANOVA results suggest that the highest yearly earnings of YouTube channels do not differ significantly based on the unemployment category, based on the data analyzed.

ii. Permutation Testing

The permutation test was conducted to evaluate whether there are statistically significant differences in the highest yearly earnings of YouTube channels across different countries. The observed difference in mean earnings was calculated as the difference between the maximum and minimum mean earnings across all countries. This value represents the largest observed disparity in earnings in the dataset. We performed the permutation test 10,000 times. In each iteration, two countries were randomly selected. The data was then subsetted to include only these countries. Within this subset, the highest yearly earnings were shuffled, and the difference in mean earnings (perm_diff[i]) was recalculated for this shuffled subset. The p-value was determined by comparing the observed difference against the distribution of differences obtained from the permutations. Specifically, it was calculated as the proportion of permutation differences that were greater than or equal to the observed difference.

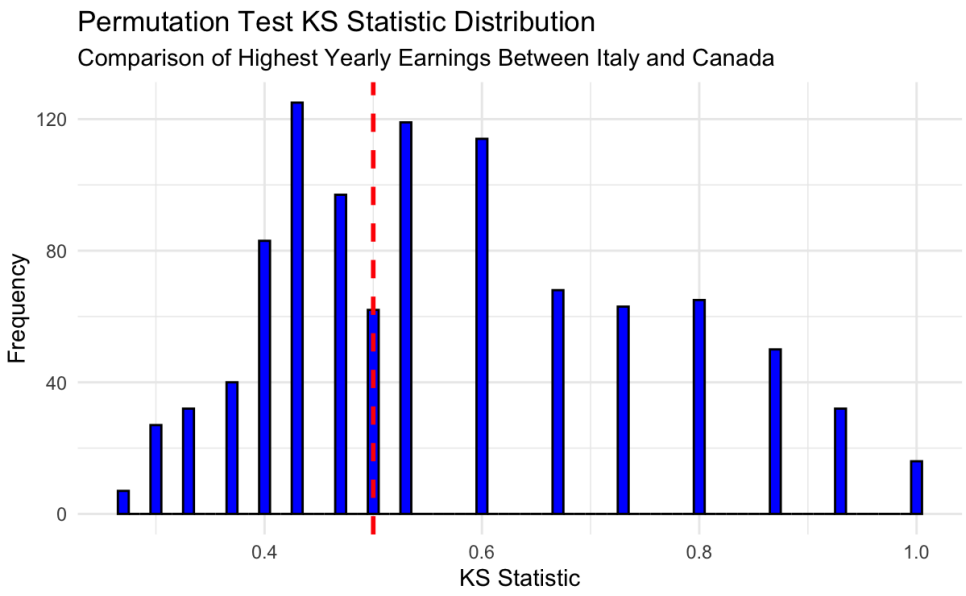
The permutation test yielded a p-value of 0.0068 (< 0.05). This indicates that only 0.68% of the permutation differences were as large as or larger than the observed difference. This suggests that the observed difference in the highest yearly earnings across countries is statistically significant. Thus, given the low p-value, we reject the null hypothesis. There is statistically significant evidence to suggest that the country of origin does impact the highest yearly earnings of YouTube channels.




A histogram was created to visualize the distribution of permutation differences. The observed difference was marked with a red dashed line for reference. The histogram shows the range and frequency of differences obtained from the permutations across all countries. The placement of the observed difference (red line) towards the extremes of this distribution visually supports the statistical finding that the observed difference is unusual compared to the permutation results.

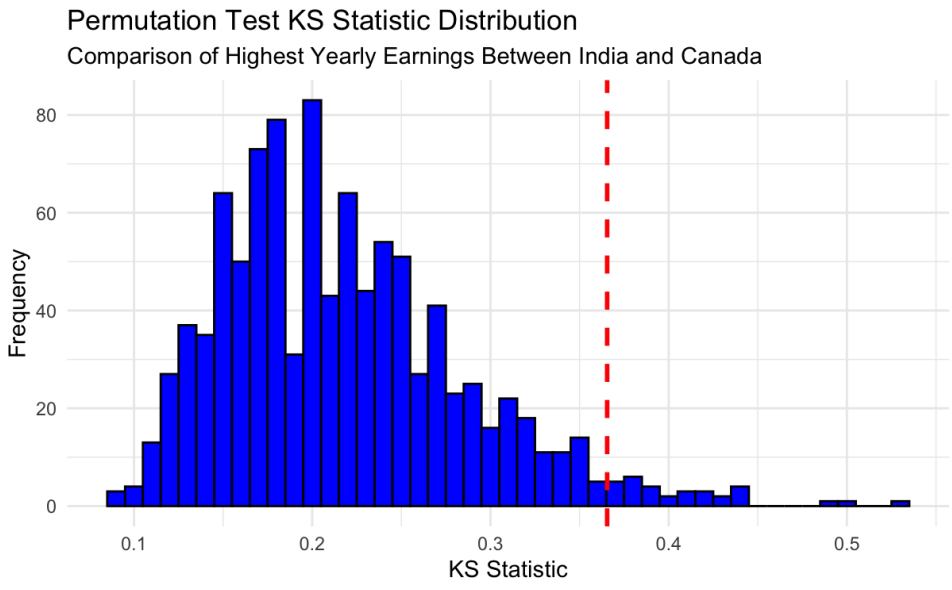
While a general permutation test across all countries can reveal overall differences in earnings, it doesn't provide insights into how specific countries compare. The KS test for specific pairs allows for a more detailed, pairwise comparison, helping to understand the relationship between two particular countries. The KS test is particularly effective in detecting differences not just in the means, but in the overall distribution of data. This means it can identify variations in the spread, skewness, or presence of outliers in earnings between two countries, offering a more comprehensive view of the differences. The KS test provides a non-parametric method to compare distributions, making it robust against non-normal data, which is often the case in real-world datasets like YouTube earnings.

We can first start by comparing the highest yearly earnings between Italy and Canada:



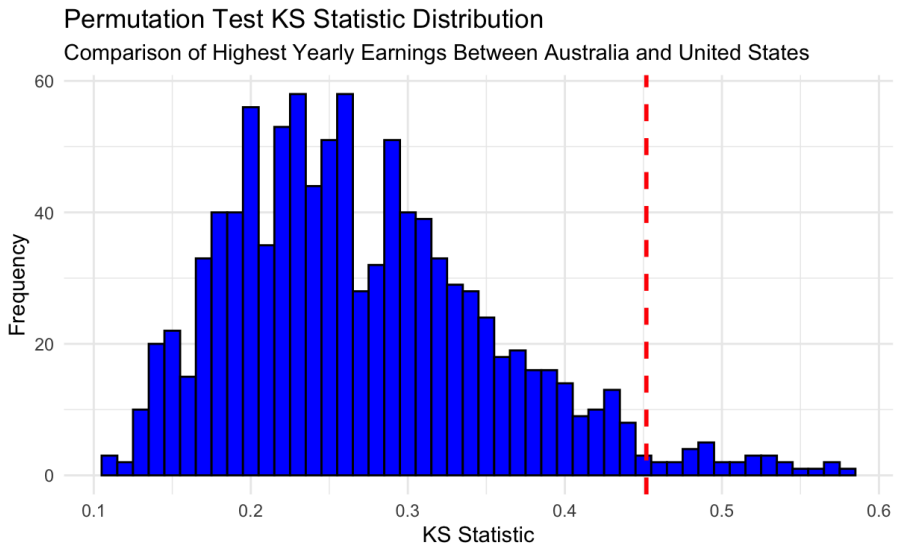

We obtained a p-value of 0.568 in this case. The high p-value indicates that there is no statistically significant difference in the earnings distributions between YouTube channels in Italy and Canada. This suggests that the earnings profiles for these two countries are quite similar.

Then we compare the highest yearly earnings between India and Canada:



We obtained a p-value of 0.031 here. The low p-value here suggests a statistically significant difference in the earnings distributions between YouTube channels in India and Canada, which implies that the earnings profiles for these countries are distinct from each other. This might reflect differences in market conditions, audience size, or monetization potential in these countries.

We might also want to see the differences in the highest yearly earnings between Australia and the United States.



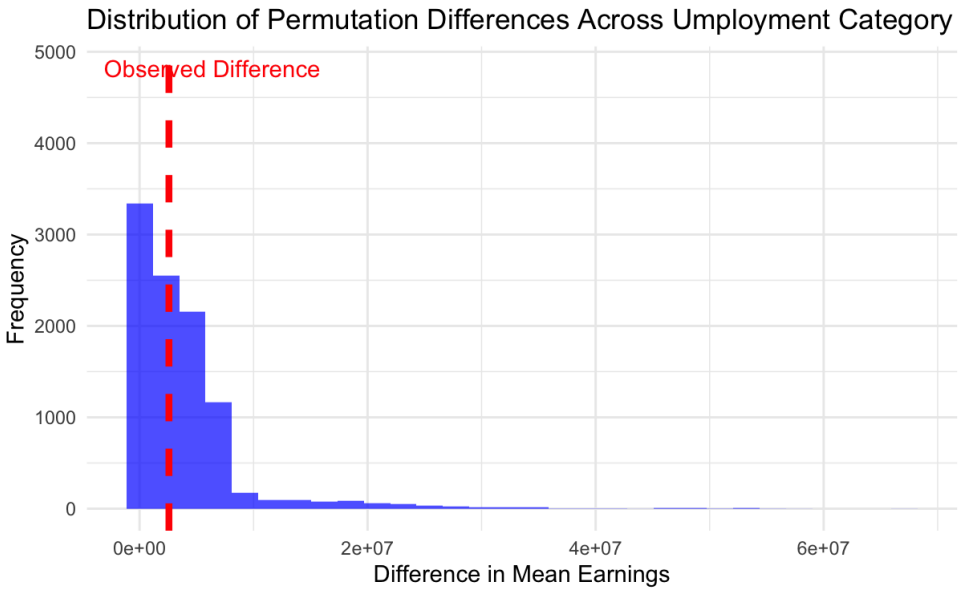

Similar to the India-Canada comparison, we get a p-value of 0.03. This low p-value indicates a significant difference in the earnings distributions between YouTube channels in Australia and the United States.

Therefore, while the general permutation test shows significant variability in earnings across all countries, the KS tests reveal that this variability is not uniform across all pairs of countries. This suggests that while there is a general global trend, specific country-to-country comparisons can yield different insights.

The differences in p-values between specific country pairs (like India-Canada vs. Italy-Canada) suggest that country-specific factors (such as market conditions, audience preferences, economic

size, or content types) significantly influence earnings. Some country pairs have similar earnings profiles, while others differ markedly. Also, the KS test is sensitive to differences in both the location and shape of the earnings distributions. This means it can detect differences not just in the averages, but also in the overall distribution patterns such as variability and skewness.

We also seek to conduct the permutation test to assess whether there are statistically significant differences in the highest yearly earnings of YouTube channels across different levels of unemployment. The p-value obtained from this test was 0.4862, which is well above the conventional alpha level of 0.05. This indicates that the observed difference in mean earnings across different unemployment categories is not statistically significant. With such a high p-value, you fail to reject the null hypothesis. The null hypothesis in this context posits that there is no significant difference in the highest yearly earnings across different unemployment categories. Based on our data, the level of unemployment does not have a statistically significant impact on the highest yearly earnings of YouTube channels.




The result suggests that the category of unemployment, used here as an indicator of broader economic conditions, does not appear to be a primary factor influencing the financial success of YouTube channels. The histogram visualizes the distribution of permutation differences after shuffling the unemployment categories - essentially simulating the null case of no category

impact. Importantly, the observed difference, shown by the red line, lies well within this distribution rather than in the tails. Only 48.62% of the permutation differences were greater than the actual observation. This suggests the observed difference could plausibly occur just by chance even if unemployment levels do not truly alter earnings.

Therefore, we conclude that countries' unemployment status does not have a statistically detectable effect on the highest yearly YouTube earnings channels located there. Variability within categories is much larger than the variability between categories.

This outcome suggests that macroeconomic factors, such as employment conditions, may not be directly influential in shaping the revenue potential of individual YouTube channels. Instead, factors specific to the content creator, such as the quality of content, audience engagement, and sponsorship deals, are likely more pivotal in determining financial success. While the geographical location of a channel does play a significant role in its earnings, finer economic indicators like unemployment rates do not seem to dictate success on YouTube. Instead, the financial outcomes for YouTube channels are more likely to be reshaped by factors related to audience interaction and the nature of the content itself, rather than by the broader economic environment, such as employment rates or population size.

## Discussion

a. Limitation

In our smoothing method, we did not find the optimal bandwidth of the LOESS smoother by hand. By default, geom_smooth() with method = "loess" selects the bandwidth using an internal algorithm, but we could potentially control it manually with the span argument. Or we could use cross-validation to select an optimal value, as discussed in our lecture. Furthermore, while there are guidelines and cross-validation methods for selecting bandwidth, the choice of bandwidth (or span) can still be somewhat subjective and may require manual tuning.

The results derived from the ANOVA and permutation tests in our study are contextual and may not universally apply to all YouTube channels or other digital platforms. It's important to note that ANOVA relies on the assumption of normal distribution within each group. This assumption

can often be violated in real-world data, such as YouTube analytics, potentially affecting the test's reliability. Additionally, ANOVA presupposes that the groups being compared are independent. This may not always be the case with YouTube channels, where inter-country influences can occur.

Permutation tests, while effective in establishing statistical significance, do not elucidate the underlying reasons or mechanisms behind observed differences. The outcomes of these tests can also be influenced by the randomness inherent in the data shuffling process. Although this can be somewhat mitigated by setting a consistent random seed, it underscores the element of chance in the results.

b. Future Research

Our research serves as a foundational exploration of how external conditions interact with YouTube channel performance. However, more detailed studies, examining how different types of content and niche categories are impacted, could reveal further insights. Investigating how broader macroeconomic indicators interplay with specific characteristics of content creators might also shed more light on the nuances influencing success metrics.

In the context of rapidly evolving digital landscapes, particularly in emerging markets experiencing increased internet access and usage, longitudinal studies tracking socioeconomic shifts and their impact on the feasibility of content creation will be highly valuable. In the future, we may choose to employing dynamic panel data models to track these changes over time. could significantly clarify how external factors correlate with YouTube revenue streams globally, providing a deeper understanding of these complex dynamics.

## Conclusion

Our research integrates various statistical methods to unravel the intricate relationship between channel earnings and popularity on YouTube, spanning across diverse content categories. Starting from looking at descriptive statistics of the merged dataset, we then utilized smoothing method to simulate success_index distribution in combat for non-linearity. Then we use bootstrap method to construct confidence intervals across youtube videos views across countries.

Noticing there appears differences among the bootstrap CIs, we placed great emphasis on permutate across countries and figure out what might be the mechanisms behind such differences. In doing permutation tests, the study underscores the pivotal influence of geographical location on a channel's potential to generate earnings. It suggests that creators based in developed countries often have a competitive edge. However, intriguingly, our analysis also points out that macroeconomic factors, such as national unemployment rates, do not notably affect the performance of individual channels.

This investigation sheds light on the inherent disparities that exist in the digital content creation, particularly in terms of economic gains. It suggests that the platform's wide reach does not necessarily equate to equitable economic benefits for all creators. To bridge this gap and ensure that content creation serves as a viable means of livelihood across diverse geographies, concerted efforts are necessary. This involves addressing the fundamental factors that contribute to these uneven barriers to monetization.

Our findings call for an enhanced empirical focus on the outcomes experienced by content creators on digital platforms, coupled with targeted policy interventions aimed at leveling the playing field. Ultimately, our study reaffirms that while geographical factors are significant, the success of YouTube channels hinges predominantly on the individual attributes of the creators, including the quality of content and the ability to engage with audiences. This observation emphasizes the dynamic and evolving nature of digital content creation, where creativity and innovation stand as the central pillars of success.

## Data Source

https://hypeauditor.com/top-youtube/

https://www.kaggle.com/code/dignil/trending-youtube-channels-titles-and-tags